様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ５ 月 ２５ 日現在

機関番号：８２６４１
研究種目：基盤研究(B)
研究期間：2011～2014
課題番号：２３３０００６３

研究課題名（和文）高次統計に基づく対話的クラスタリングの実現

研究課題名（英文）Development of Interactive Clustering based on Higher Order Statistics

研究代表者
小野田　崇（Onoda, Takashi）

一般財団法人電力中央研究所・システム技術研究所・副研究参事

研究者番号：４０３７１６６１

交付決定額（研究期間全体）：（直接経費）　　14,800,000円

研究成果の概要（和文）：近年，Webニュース，ツイッター，ソーシャルネットワークシステムなどの発展により，非常に多くの文書がネット上に氾濫している。非常に多くの文書が存在するものの，多くの文書は同じ「話題」について記述している場合がほとんどである。多くの文書の中から，この「話題」を抽出する独立潜在情報分析(Independent Semantic Analysis)を提案した。この独立潜在情報分析は，多くの文書の中から高次統計的に独立性の高い「話題」を抽出する。独立潜在情報分析により，独立性の高い「話題」が抽出できるため，その「話題」は文書要約に利用しやすく，また，各「話題」での文書のグループ化にも利用しやすい。

研究成果の概要（英文）：Recently, quite many documents have overflowed on the Internet by utilization of Web news, twitter, Social Network System, and so on. In stored documents on the Internet, many documents describe about the same “topic” and this thing is a common thing. Our research proposed the Independent Semantic Analysis (ISA) which extracts the “topic” from the quite many documents. The proposed ISA can extract independent “topics” from a high order statistical point of view. It is easy to generate an abstract by using the extracted independent “topics” and group the quite many documents by using the extracted independent “topics”.

研究分野：統計的機械学習

キーワード：クラスタリング　高次統計　知識発見　データマイニング　HAI

１．研究開始当初の背景

扱う情報の量が従来以上に増えてくることを考えると，扱う情報の構造がどのようになっているかを把握するための情報のグループ化であるクラスタリング手法の重要性が増してくる。このクラスタリング手法は大きく最短距離法などの階層的クラスタリング手法と k-means 法などの非階層的クラスタリング手法に分けられる。どちらのクラスタリング手法も基本的にはデータ間の類似性に基づいた手法であり，生成される各クラスタが類似していない，各クラスタが独立であるなどのクラスタ間の制約を陽には採用してこなかった。そのため，対話的クラスタリングにおいて，クラスタリング結果を評価するユーザはクラスタ内のデータが類似しているという視点からの評価しかできず，ユーザが与えられた結果からデータ間の must-link や cannot-link といった制約を発見し，それを計算機にフィードバックするには認知的な負荷がかかっている。一方，信号処理研究や機械学習研究の分野では，独立な信号が線形に重ね合わさった観測信号から元の独立な信号を復元する独立成分分析のような手法が提案されており，これらの手法を利用することでこれまで陽に扱われてこなかった各クラスタが独立であるといったクラスタ間の制約を取り込んだクラスタリング手法を実現でき，ユーザがデータ間の must-link や cannot-link といった制約を発見する認知的負荷を飛躍的に削減できるクラスタリング結果を提示できると期待できる。

２．研究の目的

本研究では高次統計に基づく対話的クラスタリング手法を，機械学習技術とヒューマンエージェントインタラクション技術を基に研究開発することをその研究目的とする。ここで高次統計とは，独立成分分析に代表される高次統計量に基づく，正規分布から外れた分布を対象とする統計分析技術である。この技術を利用して，データの類似性とクラスタ間の独立性を同時に満たす（右図）クラスタリング手法を研究開発する。また，クラスタリング結果をユーザに提示し，提示された結果からユーザが新たに発見したデータ間の must-link やクラスタ間の must-link という制約を満たしながら，データの類似性とクラスタ間の独立性を同時に満たすクラスタリング手法も研究開発する。さらに，データの類似性とクラスタ間の独立性を同時に満たすクラスタリング結果が人間の認知的負荷に与える影響を評価するため，開発手法を基本としたクラスタリング結果の提示とユーザ制約の指定を支援するツールを開発する。

３．研究の方法

本研究では主に，データの類似性とクラスタ間の独立性を同時に満たすクラスタリング手法の研究開発，ユーザが新たに発見した制約を満たしながら，データの類似性とクラスタ間の独立性を同時に満たす対話的クラスタリング手法の研究開発，ユーザの認知負荷の低いユーザ制約指定支援ツールの開発，および，開発手法の有用性評価の 4 つで構成される。2 種類のクラスタリング手法の研究は，[Onoda10]の研究を発展させ，独立成分分析に基づいたクラスタリング手法を開発する。ユーザの認知負荷の低いユーザ制約指定支援ツールの研究は，[Yamada10]の研究を発展させ，データの類似性とクラスタ間の独立性がユーザに伝わる効果的なインタフェースを開発する。開発手法の有用性の評価は，ベンチマークデータと被験者実験により他手法との比較を実施する。

４．研究成果

(1) 独立潜在情報分析 (Independent Semantic Analysis : ISA)の提案

文書データ $\mathbf{x}_{1\cdots N}$ は，独立な潜在情報 $\mathbf{s}_{1\cdots K}$ と，“文書での潜在情報の強度”を表す混合行列 $\mathbf{A}(\mathbf{x},\mathbf{s})$ を用いて，独立な潜在情報の線形和として次のように表すことができる．

$$\mathbf{x}_i = a(\mathbf{x}_i,\mathbf{s}_1)\cdot\mathbf{S}_1 + \cdots + a(\mathbf{x}_i,\mathbf{s}_k)\cdot\mathbf{S}_k$$

ここで，$a(\mathbf{x}_i,\mathbf{s}_1)$ は，文書データ $\mathbf{x}_i$ における独立な潜在情報 $\mathbf{s}_1$ の強度を示す値である。また，文書データ $\mathbf{x}_{1\cdots N}$ と独立な潜在情報 $\mathbf{s}_{1\cdots K}$ は単語 $\mathbf{c}_{1,\cdots,M}$ の値によって表現される。文書データを各単語 $\mathbf{c}$ が文書データ $\mathbf{x}$ の中での強さを“文書データでの単語の強度”と呼ぶ行列 $\mathbf{R}(\mathbf{x},\mathbf{c})$ による表現ができる．同様に独立な潜在情報を各単語 $\mathbf{c}$ が独立な潜在情報 $\mathbf{s}$ を特定する力を“潜在情報での単語の重要度”と呼ぶ行列 $\mathbf{V}(\mathbf{s},\mathbf{c})$ によって表す．さらに，各文書データ $\mathbf{x}$ が独立な潜在情報 $\mathbf{s}$ を特定する力を“潜在情報での文書データの重要度”と呼ぶ行列 $\mathbf{U}(\mathbf{s},\mathbf{x})$ によって表す．このとき，文書データの単語による表現と，文書データの独立な潜在情報による表現の間には，次の関係がある。

$$\sum_{\text{属性}\mathbf{c}} \mathbf{R}(\mathbf{x},\mathbf{c})\cdot\mathbf{A}(\mathbf{x},\mathbf{s}) = \sum_{\text{属性}\mathbf{c}} \mathbf{R}(\mathbf{x},\mathbf{c})\cdot\mathbf{V}(\mathbf{s},\mathbf{c})$$

重要度がその独立な潜在情報での固有の単語の組み合わせに注目する。一方で，強度は文書データ注での独立な潜在情報の組み合わせの多さを示す。

この手法では，文書データ $\mathbf{x}$ から独立な潜在情報 $\mathbf{s}$ を推定し，“文書での潜在情報の強度” $\mathbf{A}(\mathbf{x},\mathbf{s})$ に基づいて，各文書データがどの独立な潜在情報からどの程度影響を受けているのかを分析する。この手法のアルゴリズムを以下に示す。この手法で，ユーザは真の潜在情報の数 k はわからないものとする。

1. 文書データ集合 $\mathbf{X}$ を，文書データを行に，単語を列にとった行列 $\mathbf{R}(\mathbf{x},\mathbf{c})$ と

して整理する。

2. $\mathbf{R}(\mathbf{x},\mathbf{c})$ を正規化し，$\hat{\mathbf{R}}(\mathbf{x},\mathbf{c})$ を求める。
3. ステップ 2. で求めた $\hat{\mathbf{R}}(\mathbf{x},\mathbf{c})$ を次のように分解する。

$$\mathbf{U}^{\mathrm{T}} \cdot \hat{\mathbf{R}} \cdot \mathbf{V} = \mathbf{D} \Leftrightarrow \hat{\mathbf{R}} = \mathbf{U} \cdot \mathbf{D} \cdot \mathbf{V}^{\mathrm{T}}$$

ここで，$\mathbf{U}$ と $\mathbf{V}$ は独立な潜在情報での文書データと単語の重要度を示す行列である。また $\mathbf{D}$ は特異値の対角行列であり，その大きさの順に k 個の成分を抜き出し，$\mathbf{U}_k, \mathbf{D}_k, \mathbf{V}_k$ を作成する。

4. ステップ 3. で得られた $\mathbf{U}_k, \mathbf{D}_k$ を用いて，各潜在情報間の独立性が最大となるときの，“文書データにおける潜在情報の強度” $\mathbf{A}(\mathbf{x},\mathbf{s})$ を，FPICA [5] に基づいたアルゴリズムによって求める。
5. ステップ 4. で求めた“文書データにおける潜在情報の強度” $\mathbf{A}(\mathbf{x},\mathbf{s})$ の値によって，各文書データがどの潜在情報から派生しているのかを決定する。

$$\underset{\mathbf{s}}{\operatorname{argmax}}\, a(\mathbf{x}_i, \mathbf{s}_j), i \in \{1, \cdots N\},\ j \in \{1, \cdots k\}$$

(2)制約付き独立潜在情報分析(Constrained ISA : CISA)

提案した ISA の課題として，潜在情報数が未知のためユーザが指定した独立潜在情報数が適切であるかどうかが分からないことがあげられる。潜在情報の数が多い場合に，ISA で求めた潜在情報にユーザ制約を加えることで，少ない数でより独立性が高い潜在情報を求める CISA を提案した。

CISA では，ISA で求めた k 個の潜在情報にユーザ制約を加えて，$k-1$ 個のより独立な潜在情報を求めることができる。その際，次の仮定のもとで行われる。

- ISA で求めた k 個の潜在情報は，ISA で求めた $k-1$ 個の潜在情報を含んでいる。

$$\mathrm{ISA}(k-1) \in \mathrm{ISA}(k)$$

ここで $\mathrm{ISA}(k)$ は潜在情報数 k で ISA を行った潜在情報を示す。

- 残りの 1 個は $\mathrm{ISA}(k-1)$ の潜在情報のうちの 1 個から分裂して生成される。

この仮定のもとで，k 個の潜在情報から $k-1$ 個の潜在情報を求める CISA についてのアルゴリズムを次に示す。

1. 制約を加えたい潜在情報 $\mathbf{s}_i$ と $\mathbf{s}_l$ を選択する。制約を加えた新しい潜在情報 $\mathbf{ns}$ の初期値を $\mathbf{ns} = (\mathbf{s}_i + \mathbf{s}_l)/2$ とする。
2. ユーザが選択した潜在情報以外の $k-2$ 個の潜在情報を初期値として，ステップ 1. で求めた $\mathbf{ns}$ とのコサイン相関が低いものから順番に各潜在情報の独立性が最大となるように更新する。更新には FPICA [Hyvarinen97] を用いて行う。
3. $k-2$ 個の新たな潜在情報が得られたら，最後にユーザがステップ 1. で求めた $\mathbf{ns}$ の独立性が最大となるように更新する．更新には FPICA [Hyvarinen97] を用いる。

これらのステップを複数回繰り返すことで，潜在情報を 1 個ずつ減らしていくことが可能となる。

(3)結合されるべき潜在情報と適切な潜在情報数

提案した CISA の課題として，次の二つが挙げられる。

1. 独立性がより高い潜在情報の推定を行うことができる適当な 2 個の潜在情報をユーザが選択するのは難しい。
2. 潜在情報の数は最低 2 個まで減らせるが，適切である考えられる潜在情報数の数が分からない。

ここでは，この課題を克服する方法について述べる。

(3)-1 結合されるべき潜在情報

CISA を適用する仮定の下で，ユーザが選択するべき潜在情報の条件は，次の 2 つである。

- $\mathrm{ISA}(k-1)$ のときには存在しなかった新たな潜在情報
- その新たな潜在情報が生成される際に，分裂したと考えられる潜在情報

以上を満たす潜在情報を選択するために，我々は $\mathrm{ISA}(k-1)$ と $\mathrm{ISA}(k)$ とのコサイン相関を計算する。$\mathrm{ISA}(k-1) \in \mathrm{ISA}(k)$ という仮定があるので，$\mathrm{ISA}(k-1)$ と $\mathrm{ISA}(k)$ とのコサイン相関を求めると，$\mathrm{ISA}(k)$ においてある 2 個の潜在情報は，$\mathrm{ISA}(k-1)$ におけるある 1 個の潜在情報とのコサイン相関の値が大きいものとなるはずである。これは，$\mathrm{ISA}(k-1)$ における有る一つの潜在情報が，$\mathrm{ISA}(k)$ では二つに分裂したと考えるためである。このある一つの潜在情報とのコサイン相関の値が大きい二つの潜在情報をユーザが選択するべき潜在情報として提示する。

(3)-2 潜在情報数

CISA を適用する際の仮定は，ISA で求めた k 個の潜在情報は，ISA で求めた $k-1$ 個の潜在情報を含んでいること，ISA で求めた k 個の潜在情報のある二つは，ISA で求めた $k-1$ 個の潜在情報のうちのある一つから分裂して生成されているということである。この仮定が成立しなくなる直前の潜在情報数を適切な潜在情報数と考える方法を提案した。つまり，どの潜在情報間のコサイン相関が $\cos 45^\circ$ より大きい場合，仮定が崩れたとして，その直前の潜在情報数を採用する。

(4)実験結果

表 に LA Times，KOS blog と NIPS，20newsgroups のそれぞれの潜在情報数の

時の相互情報量の値を示す。表の全てのデータセットの結果を見ると，ISA によって求めた ISA($k-1$) の潜在情報の相互情報量よりも，CISA によって求めた CISA($k \rightarrow k-1$) の潜在情報の相互情報量の方が小さい値となっている事がわかる。これは ISA によって求めた潜在情報よりも CISA によって得られた潜在情報の独立性が高いことを示している。

表:ISA，CISA による潜在情報の相互情報量

	LA Times 相互情報量
ISA(8)	1.8084
CISA(9→8)	*1.7545*
ISA(7)	1.4076
CISA(8→7)	*1.3044*
ISA(6)	1.1270
CISA(7→6)	*1.0311*

	KOS blog 相互情報量
ISA(12)	3.9449
CISA(13→12)	*3.6023*
ISA(11)	3.2479
CISA(12→11)	*2.9372*
ISA(10)	2.4470
CISA(11→10)	*2.4421*

	NIPS 相互情報量
ISA(4)	0.3665
CISA(5→4)	*0.3604*
ISA(3)	0.2369
CISA(4→3)	*0.1826*
ISA(2)	0.1066
CISA(3→2)	*0.0737*

	20 newsgroups 相互情報量
ISA(10)	2.5213
CISA(11→10)	*2.2828*
ISA(9)	2.0995
CISA(10→9)	*1.9479*
ISA(8)	1.5593
CISA(9→8)	*1.5804*

<引用文献>

[Onoda10] T. Onoda, M. Sakai, S. Yamada: Careful Seeding based on Independent Component Analysis for k-means Clustering, In Proc. of 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshop on Intelligent and Web Interaction 2010, pp.112-115, 2010.

[Yamada10] M. Okabe, S. Yamada: An Interactive Tool for Constrained Clustering, World Automation Congress, IFMIP 562, 2010.

[Hyvarinen97] A. Hyvarinen, E.Oja, “A Fast Fixed-Point Algorithm for Independent Component Analysis”, Neural Computation, Vol.9, No.7, pp. 1483-1492, 1997.

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 15 件）

① Takashi Onoda, Miho Sakai, Seiji Yamada, Careful Seeding Method based on Independent Components Analysis for k-means Clustering, Journal of Emerging Technologies in Web Intelligence, 査読有, Vol.4, 2012, 51-59.

② Hiroshi Murata, Takashi Onoda, Seiji Yamada, Comparative Analysis of Relevance for SVMs-based Interactive Document Retrieval, Journal of Advanced Computational Intelligence and Intelligent Informatics, 査読有, Vol.17, No.2, 2013, 149-156.

③ 山田 誠二，水上 淳貴，岡部 正幸，インタラクティブ制約付きクラスタリングにおける制約選択を支援するインタラクションデザイン，人工知能学会論文誌，査読有，Vol.29，No.2，2013，259-267，DOI: http://dx.doi.org/10.1527/tjsai.29.259

〔学会発表〕（計 9 件）

① 西垣 貴央，小野田 崇，高次独立性に基づくクラスタリング，第 26 回人工知能学会全国大会，2012 年 6 月 15 日，山口県教育会館.

② Takashi Onoda, A Clustering Method based on Independent Component Analysis, 25th European Conference on Operational Research, July 2012, Vilnius/Lithuania.

③ 西垣 貴央，小野田 崇，データ分布の独立性に基づくクラスタリングの実験的特性分析，第 27 回人工知能学会全国大会，2013 年 6 月 5 日，富山国際会議場.

④ Masayuki Okabe, Seiji Yamada,

Uncertainty Sampling for Constrained Cluster Ensemble, In Proceedings of the 2013 Conference on Technologies and Applications of Artificial Intelligence, Nov. 2013, Taipei, Taiwan.

⑤ Takahiro Nishigaki, Takashi Onoda, A Clustering Method based on Independent Component Analysis, In Proceedings of the 2013 Conference on Technologies and Applications of Artificial Intelligence, Nov. 2013, Taipei, Taiwan.

⑥ **西垣　貴央，小野田　崇，高次独立性に基づく制約付きクラスタリング，第 28 回人工知能学会全国大会，2014 年 5 月 14 日，ひめぎんホール.**

⑦ **福永 度宗，山田 誠二，岡部 正幸，非明示的フィードバックにより訓練データ選択を支援するインタラクションデザイン，第 28 回人工知能学会全国大会，2014 年 5 月 14 日，ひめぎんホール.**

⑧ **岡部 正幸，山田 誠二，外れ値検出に基づく対話的ファイアウォールログ分析，第 28 回人工知能学会全国大会，2014 年 5 月 14 日，ひめぎんホール.**

⑨ Takashi Onoda, Constrained Independent Semantic Analysis, Conference of the International Federation of Operational Research Societies, July, 2014, Centre de Convencions Internacional de Barcelona – CCIB.

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
なし

６．研究組織

(1)研究代表者
小野田　崇（ONODA, Takashi）
電力中央研究所・システム技術研究所・副研究参事
研究者番号：40371661

(2)研究分担者
山田　誠二（YAMADA, Seiji）
国立情報学研究所・大学共同機関等の部局等・教授
研究者番号： 50220380

(3)連携研究者
岡部　正幸（OKABE, Masayuki）
豊橋技術科学大学・情報メディア基盤センター・助教
研究者番号：50362330

(4)研究協力者
西垣　貴央（NISHIGAKI, Takahiro）
東京工業大学大学院・知能システム科学専攻・博士課程